Moving from BLUP to Marker-Assisted BLUP for Genetic Evaluations

D.L. Johnson[1], C. Stricker[2], R.L. Fernando[3] and B.L. Harris[1]
[1]Livestock Improvement Corporation, Private Bag 3016, Hamilton, New Zealand.
[2]Applied Genetics Network, Obere Strasse 19, 7270 Davos Platz, Switzerland.
[3]Department of Animal Science, Iowa State University, Ames, Iowa 50011, USA.

Abstract

We discuss the development of a program for genetic evaluation based on best linear unbiased prediction using performance records, pedigree information and marker data. The covariance among relatives at the marked quantitative trait locus is calculated based on identity by descent probabilities using a descent graph sampler. The algorithm is applicable for pedigree containing ungenotyped relatives and unknown animals. Disequilibrium between markers and quantitative trait locus can be included. A variety of linear models are available and the equations are solved by preconditioned conjugate gradient methodology using iteration on data. These methods are to be applied to the Livestock Improvement breeding scheme for dairy cattle.

Introduction

In recent times molecular genetics has made it possible to partition some of the genetic variability of traits into quantitative trait loci (QTL). Since 1994, Livestock Improvement Corporation (LIC) has been involved in ventures to detect and utilise QTL in dairy cattle. Initial work was based on a granddaughter design and more recently an F2 Holstein-Friesian x Jersey crossbred trial and DNA pooling (Spelman *et al.*, 2001a,b). Marker assisted selection (MAS) has been applied within the LIC breeding scheme using a "bottom-up" approach (McKinnon and Georges, 1998), with limited success because of disappointing results with reproductive technologies used to generate full-sib families.

Three types of observable genetic loci were distinguished by Dekkers (2003). The ideal is to identify the causative mutation which has a direct effect on the trait and can then be fitted in the BLUP genetic evaluation model as a fixed effect if the genotype is observed for all individuals. Next in importance are LD markers which are in population-wide linkage disequilibrium with the functional mutation and highly likely to be in close proximity to it (1-5 cM), and are usually localised with fine mapping techniques. Finally LE markers which are in population-wide linkage equilibrium with the function mutation (LD only within family) are easily detected from the analysis of large half-sib families using sparse marker maps (20cM spacing). These three types are in increasing order of ease of detection but in reverse order for the ease and ability to utilise in selection programmes. We refer to genetic markers that have a direct effect on the trait as type I markers and the LD or LE markers that are linked to the trait as type II. Statistical methods have been developed for using marker information in BLUP genetic evaluations (e.g. Fernando and Grossman, 1989; Fernando, 2004). The linear models for genetic evaluation can be characterised by inclusion of a fixed effect if the marker has an effect on the trait means (type I or type II with LD) and a random effect to account for covariances due to cosegregation information (type II).

This paper describes developments for the implementation of a marker-assisted BLUP genetic evaluation system (MABLUP) in the LIC breeding scheme.

Materials and Methods

Models for marker-assisted BLUP

Following Fernando (2004) we assume additive gene action for the QTL linked to the marker (MQTL) and also for other loci affecting the trait (RQTL), the latter assumed to be unlinked to the markers and the MQTL.

We assume two alleles, Q_1 and Q_2 are segregating at the MQTL. If the genotypes at the MQTL are observed then the trait phenotypes can be modelled as

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{ZQ}\mu + \mathbf{Zu} + \mathbf{e} \qquad (1)$$

where **y** is the vector of phenotypic values, **β** is a vector of fixed effects, μ is the Q_2 allele substitution effect at the MQTL, **u** is the vector of additive effects of the RQTL, **e** is a vector of residuals and **X**, **Z**, and **Q** are known incidence matrices. **Q** is a column vector giving the number of Q_2 alleles for each individual. This model applies to the type I marker assuming that all individuals are genotyped.

If the genotypes at the MQTL are not observed, then **Q** is an unobservable random matrix, the elements of which depend on the information provided by observed marker loci linked to the QTL. If we define the random vector **a** with zero mean as

$$\mathbf{a} = \mathbf{Q}\mu - E(\mathbf{Q} \mid M)\mu \qquad (2)$$

where $\hat{\mathbf{Q}} = E(\mathbf{Q} \mid M)$ is the conditional expectation of **Q** given observed marker genotypes, then equation (1) can be written as

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z}\hat{\mathbf{Q}}\mu + \mathbf{Za} + \mathbf{Zu} + \mathbf{e} \qquad (3)$$

where again all incidence matrices are known. The covariance matrix of **u** is proportional to the additive relationship matrix which can be inverted efficiently (Henderson, 1976). The inverse of the covariance matrix for **a** is not sparse and cannot be inverted efficiently. However we can partition $a_i = v_i^m + v_i^p$ into the maternal and paternal MQTL alleles of individual *i,* and equation (3) now becomes

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z}\hat{\mathbf{Q}}\mu + \mathbf{Wv} + \mathbf{Zu} + \mathbf{e} \qquad (4)$$

where **W** is an incidence matrix and now the inverse of the covariance matrix of **v** can now be inverted efficiently (Wang *et al*., 1995).

Comparing equations (1) and (4), the different types of genetic loci can be associated with the terms involving μ and **v**. Under equilibrium, the matrix $\hat{\mathbf{Q}}$ is constant and so $\mathbf{Z}\hat{\mathbf{Q}}\mu$ can be dropped from equation (4). In this situation only cosegregation information will contribute to the analysis through covariances among MQTL effects. On the other hand, when disequilibrium is complete and all marker genotypes are observed, $\hat{\mathbf{Q}} = \mathbf{Q}$ and **v** is null so that only disequilibrium information and no cosegregation information contributes to the analysis, and equation (4) reduces to equation (1). When disequilibrium is partial, equation (4) applies where disequilibrium information will contribute to the mean of MQTL effects and cosegregation information will contribute through covariances between MQTL effects.

For type I markers, when not all individuals are genotyped, some elements of **Q** are no longer observed but can be replaced by their conditional expectations given observed genotypes. A model similar to equation (3) can be applied where the random vector **a** corresponds to animals with missing genotypes. An inverse for var(**a**) can be computed efficiently (Fernando, 2004).

Mean of MQTL additive genetic values

Each element of the matrix **Q** is the sum of two Bernoulli variables so that the conditional expectation has elements

$$\hat{Q}_i = p_i^m + p_i^p$$

which is the sum of the probabilities that the maternal and paternal MQTL allele states for individual *i* are identical to allele Q_2 given marker information. Each of these probabilities in turn can be expressed as the probability that the maternal (paternal) MQTL allele originated in a given founder haplotype times the probability that the founder haplotype has MQTL allele Q_2, summed over the haplotype states. The latter probabilities are the disequilibrium parameters.

Covariance of MQTL additive values

The variance of the maternal MQTL effect for individual *i* is $\mu^2 p_i^m (1 - p_i^m)$ and similarly for the paternal allele and therefore depend on marker genotypes under disequilibrium. Covariances between MQTL effects are based on the probabilities of the descending QTLs (PDQs) and can be calculated using a recurrence formula which is independent of the level of disequilibrium. The PDQs are estimated by means of a descent graph sampler (Schelling, 2004; Stricker *et al.*, 2002). Allelic origin is sampled at multiple marker loci given an arbitrary pedigree and known marker information. Disequilibrium between markers and QTL can be accounted for through extending the allelic origin samplers to estimate founder haplotype origin probabilities. For each individual four non-zero descent probabilities are derived according to gamete identification by parental origin. For the paternal allele of the individual we have the probabilities of inheriting the paternal or maternal MQTL allele of the sire conditional on the observed marker information, and similarly the probabilities that the maternal allele in the individual was descended from the paternal or maternal MQTL allele of the dam. There are only two independent probabilities per individual because the paternal and maternal origin probabilities must sum to unity. The elements of $\text{var}(\mathbf{v})^{-1}$ are then derived using simple tabular rules based on knowledge of the PDQs (Wang *et al.*, 1995; B.L. Harris and D.L. Johnson, unpublished). The algorithm eliminates singularity problems by including only independent MQTL effects (i.e. when any of the PDQ probabilities is unity), provides a non-singular inverse for any pedigree, and a method to prune MQTL effects not linked to animals with data. Inbreeding at the QTL locus can also be incorporated.

Software development

The MAS/GAS project which we describe has focused on two main modules: (i) calculation of the PDQs, and (ii) the development of MABLUP for solving a variety of linear mixed models.

The MABLUP module, for both single- and multiple-trait analyses, is designed to handle repeated measures, maternal effects, multiple QTL in equilibrium or disequilibrium with flanking markers. MQTL effects are assumed to be additive in all cases. The situation of pleiotropy (effect of QTL on several traits) can be handled for two situations. If we assume two alleles at the QTL then allelic effects at subsequent traits can be modelled as a linear combination of the allelic effects at the first trait (the scale model). Alternatively, the covariance matrix at the MQTL can still be obtained as a direct product between **G** and the covariance matrix among the MQTL effects for the different traits. But under disequilibrium only the scale model is available as the expression as a direct product is not possible because the MQTL covariances among traits are now dependent on the MQTL means. The mixed model equations are solved using preconditioned conjugate gradient methodology and iteration on data for large models.

Discussion

The challenges in the future, in addition to further theoretical developments in the multiple trait case under disequilibrium, involve testing the package on large data sets (of the order of tens of thousands) and the estimation of parameters required for the model. The latter include development of an approach to maximise the likelihood for estimation of disequilibrium parameters and variance components due to the MQTL and polygenes. Other issues include the reduction in size of the equations through elimination of missing MQTL observations.

Current animals and ancestors involve some 15,000 animals in the LIC breeding scheme. Testing on this size of data will undoubtedly lead to further efficiencies in the software. Integration of MABLUP in the breeding scheme raises issues as to which animals to genotype. The source of bull mothers in the New Zealand population comes from the commercial cow population and so genotype information is not readily available as would be the case for a central nucleus. The flow of genes from the commercial tier may change in the future due to increased genetic

progress (therefore increased lag between tiers) or through limitations in trait recording and genotyping in the wider population. Thus the national herd as a source of bull mothers may be less important in the future and a subset of elite cows will likely emerge. The use of reproductive technologies among contracted cows leads essentially to a dispersed nucleus system (Meuwissen, 1991). We need to identify those cow families that contribute most to the LIC breeding scheme and give consideration to factors such as cost, technical feasibility and rates of genetic improvement.

Successful implementation of MAS requires an integrated approach involving economic aspects, business goals and risks and markets. Economic values for a selection index including MQTL information should encapsulate value on farm, marketing criteria and time to fixation of favourable alleles.

Enhancements to the LIC breeding scheme over time has led to increased rates of genetic gain. New technologies such as genetic markers have promise of further improvements. To facilitate an improved breeding scheme we require systems modelling of those components of the dairy industry that influence genetic gain and productivity.

References

Dekkers, J.C.M. 2003. Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. *Proc. 54th EAAP,* Rome, Italy.

Fernando, R.L. & Grossman, M. 1989. Marker assisted selection using best linear unbiased prediction. *Genet. Sel. Evol. 21,* 467-477.

Fernando, R.L. 2004. Incorporating markers into genetic evaluation. *Proc. 55th EAAP,* Bled, Slovenia.

Henderson, C.R. 1976. A simple method for computing the inverse of a numerator relationship matrix used in predicting breeding values. *Biometrics 32,* 69-83.

McKinnon, M.J. & Georges, M. 1998. A bottom-up approach to marker assisted selection. *Livest. Prod. Sci. 54,* 229-250.

Meuwissen, T.H.E. 1991. The use of increased female reproductive rates in dairy cattle breeding schemes. *Anim. Prod. 52*, 21-31.

Schelling, M. 2004. Deterministic calculation and stochastic simulation in multi-point linkage analysis. *Ph.D. thesis, Swiss Federal Institute of Tecnology ETH No. 15335.*

Spelman, R.J., Coppieters, W., Grisart, B., Blott, S. & Georges, M. 2001a. Review of QTL mapping in the New Zealand and Dutch dairy cattle populations. *Proc. 14th AAABG,* 11-16.

Spelman, R.J., Miller, F.M., Hooper, J.D., Thielen, M. & Garrick, D.J. 2001b. Experimental design for QTL trial involving New Zealand Friesian and Jersey breeds. *Proc 14th AAABG,* 393-396.

Stricker, C., Schelling, M., Du, F., Hoeschele, I., Fernandez, S.A. & Fernando, R.L. 2002. A comparison of efficient genotype samplers for complex pedigrees and multiple linked loci. *Proc. 7th WCGALP, Montpellier, France,* 21-12.

Wang, T., Fernando, R.L., van der Beek, S., Grossman, M. & van Arendonk, J.A.M. 1995. Covariance between relatives for a marked quantitative trait locus. *Genet. Sel. Evol. 27,* 251-274.